Computational Learning Theory

Principle Investigators and Collaborators

Clint Scovel, PI	CIC-3	Don Hush, co-PI	CIC-3
Madhav Marathe, co-PI	TSA-2	Chris Barrett	TSA-2
Christian Reidys	TSA-2	James Howse	X-8
Mark Ettinger	NIS-8	Simon Perkins	NIS-2
Gabriel Istrate	CIC-3/CNLS	Leonid Gurvits	IAS
Adam Cannon	Columbia	Vladimir Koltchinskii	UNM
Richard Stearns	SUNY Albany		

Point-of-contact: Clint Scovel, CIC-3, Phone: 665-4721, Email: jcs@lanl.gov
Funding Request: $800K per year for 4 years.

Executive Summary

The use of empirical observations to augment first principles knowledge in the development of a scientific model is referred to as *learning*. The complexity of many learning problems of national importance extends far beyond the capabilities of current scientific methodology. Specifically we are often unable to compute acceptable solutions and simultaneously assess their accuracy. Without doing so, proposed solutions may generate disastrous consequences. Computational Learning Theory (COLT) defines a formal mathematical framework that directly addresses these issues by facilitating the development of efficient computational solutions while simultaneously assessing their accuracy. In recent years COLT has revolutionized the solution of classic problems in classification, prediction and regression, and has established itself as the preferred scientific framework in which to attack these problems.

Several important application domains at the laboratory, including non-proliferation and discrete simulation based decision making for infrastructure analysis, require a resolution of learning issues related to adaptation, optimal search, and game theory. We propose to study foundational questions in Computational Learning Theory (COLT) with the goal of strengthening two of the laboratory's core competency areas (theory modeling and high performance computing, and analysis and assessment), with relevance to others as well (see `http://www.lanl.gov/worldview/science/core/`). Specifically, we will focus on the following inter-related topics:

- the development of mathematical tools for attaining guarantees on accuracy,
- the study of the computational issues concerning specific learning problems and the design and analysis of new and provably efficient algorithms for them,
- a formal study of the interplay between computational resources and accuracy, leading to new techniques that can address previously unsolved problems,
- extension of the basic learning theory, in particular develop a theory of learning in games.
- the illustration of our research by demonstrating its use in analysis of infrastructure simulations and non-proliferation.

Specialist Reviewers: Emanuel Knill (CIC-3), Larry Winter (CIC-3), David Campbell (Dept. of Physics, UIUC), David Wolpert (NASA), Umesh Vazirani (U.C. Berkeley)

1 Introduction

LANL has long been the leader in the use of high performance computers to perform the massive amounts of numerical calculations required to solve real physical problems. Increasing computational capabilities have allowed the discretization of first principle models to be refined to the point where inaccuracies in the model may exceed inaccuracies due to discretization. As an alternative to such refinement it is becoming increasingly popular to use the computer to *learn* a more accurate model from *experimental data*. In recent years this paradigm has evolved into a mature science that has revolutionized the modeling process in many scientific disciplines. The use of the computer to learn models from experimental data often involves deep computational issues at the cutting edge of computer science research. Furthermore, from a practical standpoint, it forms the basis of a number of extremely important laboratory projects. Examples include the design of nuclear weapons without testing, understanding the human genome, the nonproliferation of nuclear weapons, the detection of hard and deeply buried targets, the preservation of physical and information security, the design and development of simulations for complex socio-technical problems and the efficient utilization of massively parallel computing resources. All of the above problems share a common framework in that the computer is used to design and implement models, based on empirical information, that can be used to make decisions and/or predictions that have direct bearing on the problem solution. We refer to this general problem domain as *learning*, and it includes many domains from classical statistics such as classification, regression, prediction, estimation, decision theory, optimal design, and game theory. The fundamental resources available to design and build these models include: first principles knowledge, empirical information, and computing resources. By their very nature these resources are scarce, and thus we are forced to solve these problems with incomplete first principles knowledge, limited empirical information, and finite computing resources. For example, in the computer security problem our first principles knowledge falls far short of providing a complete model for a computer hacker. In addition, the available empirical information (e.g. positive and negative examples of computer intrusions) is also insufficient to completely characterize the computer hacker. Similar statements can be made for the other problems above.

The consequences of incomplete first principles knowledge and/or empirical information can be severe. For example consider a predictive model of computer intrusions designed from positive and negative examples. If we have sufficient first principles knowledge to indicate that the correct model separates the data by some hyperplane, then the problem of finding a separating hyperplane is computationally tractable. But the accuracy of such a hyperplane (i.e. its performance on future data) must be assessed. While classical statistics provides guarantees on accuracy as the sample size goes to infinity (corresponding to complete empirical information), the practitioner is concerned with guarantees based on finite (possible small) sample sizes (incomplete empirical information). The need for accuracy guarantees based on a finite number of samples is ubiquitous. The celebrated Vapnik-Chervonenkis (VC) theory of generalization provides such guarantees, and consequently forms a cornerstone in the field of Computational Learning Theory. Now suppose that our first principles knowledge is insufficient in that the hyperplane model is incorrect. Then the problem of finding a hyperplane that minimizes empirical error is computationally intractable. Indeed examples of how insufficient first principles knowledge can lead to an explosion in the computational resources required are pervasive. The recent field of *Computational Learning Theory* (COLT)[1] has provided a resolution to this historical dilemma by posing a computationally

[1]The foundations of COLT were laid in the two seminal papers by Valiant in 1984, where he introduced the probably approximately correct (PAC) model of learning.

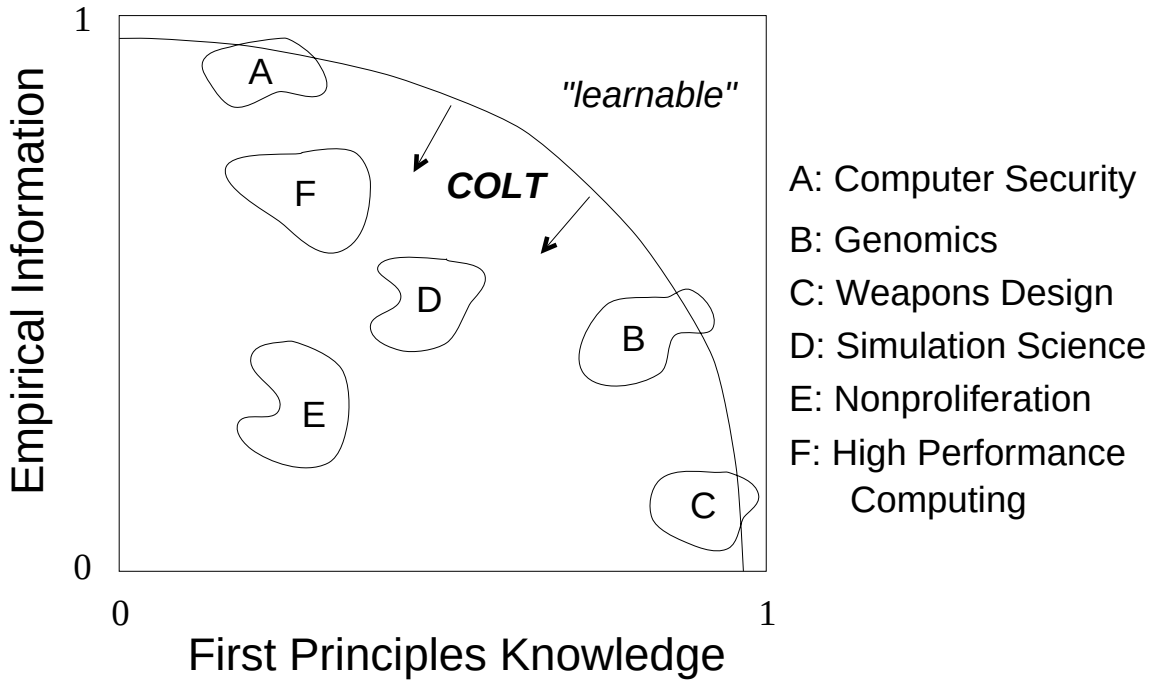

Figure 1: Current problems at LANL and their relationship to the learnability boundary.

tractable problem for hyperplane design whose solution possesses guarantees on accuracy. This was accomplished in part by extending the VC theory. Consequently the field of Computational Learning Theory has advanced linear classifier design from a severely inadequate state to one that satisfies the key concerns of the practitioner; i.e. *with reasonable computational resources he can produce a solution with accuracy guarantees, and furthermore he understands the trade-off between computation and accuracy.*

COLT defines a formal mathematical framework that enables a rigorous analysis of *both* accuracy and computational resources, and the interplay between them. This simultaneous attention to both issues is essential to the successful solution of complex learning problems. The diagram in Figure 1 is used to illustrate these concepts. The horizontal axis represents the fraction of first principles knowledge where 1 indicates the availability of complete first principles knowledge, and the vertical axis represents the fraction of empirical information available. We have highlighted regions where important problems at LANL might lie in this diagram. We also show a curve that partitions the diagram in such a way that the region to the upper right corresponds to those problems for which we can guarantee that a suitably accurate solution can be produced with reasonable computational resources (we call these "learnable" problems). COLT has already substantially increased the size of this region. The goal of this proposal is to significantly improve our understanding in this field and thereby provide sound mathematical and computational solutions for a number of important problems encountered LANL.

2 Background

Although the problem of building models from empirical data can be traced back to (at least) Gauss, this field entered a new era with the introduction of the *Probably Approximately Correct* (PAC) learning model by Valiant in 1984. This model provided a more rigorous framework in

which to study this problem. This enrichment was due in large part to the integration of Statistical Learning Theory (the design of predictors from sample data) and the Theory of Computation. The time for integration was ripe. Advances in statistical learning theory (e.g. the work of Vapnik and Chervonenkis) made it possible to provide theoretical accuracy guarantees that, for the first time, were based on a *finite* number of training samples (i.e. the guarantees were *non-asymptotic*). Consequently, this made it possible to incorporate the computational complexity framework developed by the computer science community to account for computational issues concerned with producing the model. The resulting PAC model provided the first rigorous framework that addressed the concerns of the real-world practitioner, and became the first successful model in the COLT framework.

Since COLT accounts for these quantities without imposing restrictions on how they are integrated, an unprecedented degree of exploration and development has occurred, elevating model design to a whole new level. Some of these developments are surprising in that they account for these quantities through unconventional means. For example Vapnik's Support Vector Machines defy conventional wisdom by introducing a mechanism for controlling performance in a high dimensional space [25]. Freund and Schapire's Boosting algorithm dramatically simplifies classifier design by combining mediocre classifiers in an elementary way to achieve near optimal performance [13]. Breiman's innovative Bagging technique combines models built on resamplings of the data to yield a superior model [6]. These new methods have performed extremely well in practice, but are just the beginning. Indeed, the goal of this proposal is to further the development of such methods and to extend COLT to successfully address complex problems of national importance such as those mentioned in the introduction.

Organizations such as AT&T, IBM and MicroSoft are facing learning problems that are similar in size and complexity to those at LANL and consider COLT to be critical to their success. Consequently they have substantial research efforts in this area. Pursuing research in COLT will place LANL among select academic and industrial research groups and will impact LANL projects that are aligned with the strategic plans of all three directorates (see `http://www.lanl.gov/labview/org/planning/docs.html`). Two important programs that will directly benefit from this research are: (a) modeling and simulation projects being undertaken in TSA especially those involving game theoretic settings and (b) detection and characterization of hard and deeply buried targets such as hostile nuclear and biological weapons facilities. Other specific laboratory benefits include: (i) new program development avenues with DOD, DOE, DOT (See TSA webpage for more information) as well as commercial entities including wireless and electric power industry, (ii) rigorous mathematical foundations for making predictions that will be crucial for maintaining competencies in basic thrust areas at the laboratory including Stockpile Stewardship, Non-proliferation, Simulation Science, Bioscience and Bio-technology, and Information Security, (iii) problem domains and research areas that will be crucial in attracting/recruiting people with skills in critical areas of computer science and discrete mathematics (iv) unification of algorithmic discrete mathematics and machine learning at LANL, (v) new algorithms and probabilistic tools applicable to other applications at LANL.

3 Proposal

The proposed research consists of four basic goals (i) the study of generalization error (accuracy) guarantees, and the trade-off between accuracy guarantees and computational resources, (ii) the development and analysis of efficient algorithms and associated lower bounds (from a computational resource standpoint) for solving learning problems, (iii) extensions of the basic COLT paradigm, in particular the study of learning in games and (iv) the illustration of our results by demonstrating

their use in three important and deliberately chosen application areas

3.1 Generalization Error and tradeoffs with Computational Resources

For many learning problems the measure of accuracy is called the generalization error, and is defined to be the average error with respect to the distribution generating empirical observations. Generalization theory is concerned with specific learning mechanisms that can be used to control generalization error. For example, the Vapnik-Chervonenkis theory guarantees control of generalization error for learning mechanisms that minimize the training error. More recently Bartlett [5] has shown that a different mechanism called *margin* can be used to control generalization error, and in many problems defeats the curse of dimensionality. However, there are many important learning mechanisms which have no guarantees on generalization error. Examples include Maximum Likelihood Estimation, Boosting, Bagging, and Stacking. In addition, the commonly used technique of cross-validation does not provide rigorous guarantees on generalization error. We propose to resolve these issues on two fronts. The first is the development of mathematical tools that form the basic components of the theory of generalization. For example the theory of concentration of measure, with outstanding results such as Hoeffding's inequality and Talagrand's inequality, is essential to any theory of generalization. We propose to extend this theory by developing inequalities that are relevant to large classes of learning mechanisms. The second is the development of generalization bounds for specific learning mechanisms like those mentioned above.

One of the most exciting new avenues of research that has emerged from the COLT framework is the study of the trade-off between generalization error and computational resources achieved through variation of the learning mechanism. VC theory has motivated the choice of empirical error minimization as a learning mechanism because it possesses guarantees on generalization error. However this mechanism often leads to a computationally intractable optimization problem. For example, consider the landmark problem of designing a linear classifier with optimal generalization error. VC theory shows that a linear classifier that minimizes empirical error possesses guarantees on generalization error, but determining such a classifier is computationally intractable. Recently this issue has been resolved through the introduction of an alternate learning mechanism called *margin* which forms the foundation of Vapnik's Support Vector Machines [25]. Support Vector Machines simultaneously determine a computationally tractable optimization problem while guaranteeing the generalization error of its solution [25]. *This landmark technique represents the first time the practitioner can be assured to obtain a solution with guarantees on generalization error using a reasonable amount of computational resources.* It also enables the first *quantitative* study of the trade-off between generalization error and computational resources. We propose two studies of this trade-off: a quantitative analysis of the trade-off for Support Vector Machines, and an investigation into new learning mechanisms for which the trade-off can be quantified.

3.2 Learning Algorithms and Computational Complexity

The COLT framework requires that the learning process be computationally efficient. Typically a learning mechanism can be translated into an optimization problem whose criterion is a function of the empirical data. Characterizing the intrinsic computational complexity of this optimization problem (i.e. determining whether the problem is computationally tractable) is an essential first step. This not only determines a coarse bound on the computational resources required to solve the problem, but often identifies its computationally difficult components, and points towards an appropriate classes of solution techniques. However, many popular learning algorithms have been

developed without concern for computational resources. In fact, many do not have a proof of finite time convergence (i.e. the algorithm has not been shown to produce a suitable finite time solution for all inputs and all initial conditions). For example, the Pocket algorithm is considered to be one of the most efficient algorithms for linear classifier design, and although it is known to converge asymptotically, it has no suitable finite time stopping condition and therefore no reasonable bound on its run time [21]. This is typical of existing learning algorithms. We propose work on two fronts:

1. the analysis of existing learning algorithms (i.e. the development of tight bounds on computational resources), and
2. the development of new learning algorithms with improved efficiency.

The Analysis of Existing Learning Algorithms

A recent hot topic in Theoretical Computer Science has been the study of the convergence time of ergodic Markov chains. Such Markov chains are used for randomly sampling and approximately counting combinatorial objects, and these methods have led to significant improvements in the analysis of algorithms from Statistical Mechanics. A relatively new development, *perfect sampling using coupling from the past* enables the development of algorithms that can eliminate the initial transient bias, by sampling *exactly* from the stationary distribution, without a too severe computational overhead. There is more than a superficial analogy between this group of ideas and learning theory. In fact one can regard, in a fairly naive way, a particular learning algorithm as Markov chain having as states the (usually infinite) set of hypotheses, and transitions determined by the addition of new training samples. Therefore, we hope that geometric techniques that have been so useful in the analysis of Markov chains can provide significant insights when applied to learning algorithms.

Techniques for designing efficient learning algorithms

In recent years the COLT community has determined that many of the learning mechanisms that are most useful in controlling generalization error do not lead directly to optimization problems that admit computationally efficient learning algorithms. Indeed, in many learning problems the intrinsic computational complexity is *hard* in the first place, in that it admits no algorithm whose worst case run time is polynomial for all problem instances. A similar problem existed for many classic problems in Computer Science (e.g. Traveling Salesman (TSP), KNAPSACK, PARTITION), but has been successfully approached through the use of two technical tools: the *probabilistic method* championed by Erdös, and ideas from *parameterized complexity*.

The probabilistic method has revolutionized the study of landmark problems in Computer Science by providing resolutions to issues which have eluded researchers for several decades. For example it has shown that while the worst case complexity of HAMILTONIAN CYCLE is exponential, its typical case can be solved in polynomial time. Exactly when this can be expected to hold is the subject of vigorous inquiry. There is little doubt that similar results exist for learning problems. In fact the theory of generalization has developed and used tools, such as the theory of the concentration of measure, which are fundamental to the probabilistic method. Surely such tools will be useful in the application of the probabilistic method to learning problems. We propose to apply and extend the use of the *probabilistic method* to the analysis of learning algorithms. The second approach, parameterized complexity, deals with the apparent intractability of interesting combinatorial problems by recognizing that instances of these problems have natural "parameters", that are "small" in most practical instances, and develops algorithmic methods that work efficiently (both theoretically and in practice) for such "small" parameter values. An illuminating

example is the register allocation problem for the compilation of structured imperative languages. In general this amounts to a graph-coloring problem, that is NP-complete. Nevertheless, it has been shown that graphs arising from structured programs that arise from many programming languages are characterized by a small value of a parameter called *treewidth*, which enables the development of efficient algorithms for this problem. Parameterized Complexity is a concept that has already been shown to capture the complexity of central quantities from Computational Learning Theory (such as the Vapnik-Chervonenkis dimension). We propose to incorporate such ideas into learning algorithms as they seem to be natural and promising.

3.3 Extensions of Learning Theory

There have been several recent extensions to the basic COLT framework for prediction and classification. For example extensions have been considered that treat non-independent data and loss functions other than generalization error. However, extensions that consider all of the relevant aspects of large complex computational problems have yet to be explored. For example, a preliminary study of the code validation problem of stockpile stewardship indicates that the COLT framework can (and should) be applied. However, it will require extensions such as accuracy guarantees in terms of relevant distributional assumptions and the extension of the VC theory to the predictions defined by numerical integration algorithms for partial differential equations. In the next section we provide a more detailed description of the development of an extension of COLT to an important problem at LANL.

Game Theoretic Settings for Learning

The setup of the *theory of learning in games* [10] is as follows: agents from a large population interact by repeatedly playing a fixed game and updating their behavior based on the outcomes. Particular models are obtained by specifying the game, the interaction topology and the update mechanism. The main thrust of the theory is to explain the emergence of various types of game-theoretic equilibria as the outcome of an evolutionary process (rather than seeing them as steady-state properties). Such models are of interest in a wide range of areas, from Biology to Computer Networks. Of special interest to the Laboratory's mission is the work by Aumann, Maschler and Stearns [1], who have developed such ideas in the context of studying the dynamics of arms control negotiation.

This setting corresponds to a extension to a distributed version of Valiant's PAC-model. The major difficulty is not related to computational intractability of each local step, but with the large number of agents and the local nature of interactions. Some important issues include: (i) computational power and efficiency and (ii) role of information (See Fudenberg and Levine [10] for more discussion on this subject.) Specifically we are interested in understanding (i) how does the computational power of the players, as well as their interaction structure, affect their learning process? (ii) given a simulated environment, how efficiently can one reach a "quasi-equilibrium" state without mimicking the real world learning process, (iii) given the impossibility of each player knowing the state of the entire system, are there efficient ways to collect and feedback such information and (iv) what effect does summary information have on the overall system dynamics? We propose to extend the basic PAC-learning framework to incorporate these issues.

The above discussion is relevant to new and emerging areas of Distributed Artificial Intelligence and economic mechanisms in computing. Game-theoretic techniques, such as *mechanism design* start from a specification of the agents' preferences and design the interaction process characteristic to the "learning in games" setup to attain a "socially desirable" equilibrium. Such mechanisms

have been recently studied from the standpoint of algorithmic efficiency [23]. They assume, however, rather detailed knowledge (in the form of a utility function) of the agents' preferences. Such information is often not available beforehand, yet we have to develop mechanisms that are robust enough to encompass the "real world" applications. We propose to extend the COLT framework to the simulation environment to tackle this problem.

We have conducted preliminary investigations, [11] and [19], mainly in regards to factors (such as the interaction topology and temporal structure) that affect the success of learning. We have also considered combinatorial and computational complexity questions related to this topic in [2]. We propose to further study these issues to better understand this promising new field.

3.4 Applications

Computer Security

An important problem in computer security that can be addressed by the COLT framework is *intrusion detection*, i.e. the detection of unauthorized network activity. It can be formulated as a learning problem as follows. A fundamental unit of computer activity is a *connection* between a user and a computational resource. Information for representing a connection can be obtained from the communication packets used in its establishment. Once a definition of an intrusion has been articulated empirical observations (i.e. "connections") may be generated and labeled as intrusive or not by security experts. DARPA has done this in its 1998 Intrusion Detection Evaluation Program which collected data from a local area network simulating a typical U.S. Air Force computing environment. In our recent case study with this data the detectors produced by Support Vector Machines were superior in many ways to a comprehensive collection of 10 of the most popular detectors available today. In particular they provided the best combination of computational efficiency and guaranteed accuracy. Their accuracy was second best overall and very close to the best (.14% error compared to .11% error), and among the top performers the computational resources required by Support Vector Machines were at least 100 times better (15 seconds compared to a range from .5 to 3 hours). Its important to note that the abundance of data in this particular problem made it possible to determine the accuracy of all of the methods. However without such an abundance the better performer would never have been discovered and the only detector whose accuracy would have been known with high confidence was the Support Vector Machine.

These results strongly suggest that the intrusion detection problem at LANL could benefit significantly from the the application of similar techniques. The current intrusion detection system at LANL (NADIR) was developed in much the same way as most intrusion detection systems in that it was designed without the use of ground truth empirical data. The accuracy of systems developed in this way is unknown, and in fact could be worse than random guessing. The employment of a substandard detector leads to intrusion rates that are significantly higher than necessary and therefore may severely compromise computer security at LANL. A major advantage to knowing the accuracy of a set of detectors is the ability to choose the best.

To track the dynamic nature of intrusions current systems must be updated each time a new intrusion scheme is discovered. A similar situation exists for virus detectors on personal computers. Each time a new virus is discovered (usually by someone who has already suffered its consequences) a new detector must be designed and installed on each individual system. A detector that learned from such examples can adapt to the emergence of new viruses without being redesigned and reinstalled.

In summary, COLT provides a framework in which to determine high performance detectors which can be updated with much less human interaction.

Non-proliferation

An important non-proliferation problem is the detection and characterization of hard and deeply buried targets such as hostile nuclear and biological weapons facilities. Remotely sensed data (e.g. seismic, satellite imagery, infrared imagery, and SAR imagery) may contain information that can be used to to detect such facilities. Specifically, consider the problem of detecting the construction of a nuclear weapons facility using seismic sensor data. In a first principles approach a seismologist and a nuclear weapons plant designer would work together to develop a detector based on nuclear facility construction processes and their corresponding seismic disturbances. Incomplete first principles knowledge motivates the use of COLT for detector design. Empirical observations consist of seismic data labeled by experts according to whether the data represents a hard and deeply buried target or not. Equipped with such data numerous techniques are available for designing such a detector, but only those that satisfy the COLT framework are guaranteed to produce a detector whose accuracy is known within well-defined bounds, and do so with a reasonable amount of computing resources. Providing guaranteed performance bounds in this highly uncertain problem domain would fulfill an important national security need.

Simulation and Modeling

Simulations of large scale socio-technical systems often involve intricate learning and game theory. Important projects undertaken at the laboratory in this spirit include: ELISIM (simulation system for understanding the deregulated power market), TRANSIMS (Transportation simulation and analysis tool) and Jointsim (Military applications). For example we discuss TRANSIMS in detail. We use a feedback mechanism between the route finder and the microsimulation as a way to find "relaxed" routes. The feedback mechanism provides a way to spread the information among the players (drivers) who we assume would like to settle down to routes that come close to minimizing their utility function. Each traveler has incomplete information about the system. This remains true even in a simulated environment where it is computationally inefficient to make available such information. Now consider a feedback step in which each traveler will decide whether or not to modify his/her route. Such a step depends on how/what data is fed back. For certain types of summary information, a driver might deem his route as sub-optimal and thus like to change it; with other type of information, the current route might be adequate. Two important issues that are present in this example are: (i) quality of routes is dependent on the feedback information and (ii) theoretical constructs that work with complete information and state space that keeps track of all the travelers is not computationally feasible. As pointed out in Section 3.3, addressing this problem requires an extension of the basic learning paradigm. We propose to study such extensions. Specifically, we believe that it is possible to build a model that incorporates the theory of repeated games with incomplete information [1] and the PAC-learning model [24] for understanding problems discussed above. Similar issues arise in developing simulations to understand the deregulated electrical power utility. Here again, the players seek bidding strategies (for buying or selling power at a specified price) and utilizing incomplete information. The problem is extremely important: deregulation of electrical power industry is creating an environment where such strategies will have a profound effect on the stability and the security of the underlying system.

4 Schedule and Deliverables

Because understanding the generalization error of learning algorithms and the computational complexity of learning algorithms is so fundamental to all aspects of COLT, they will be addressed at a fundamental level throughout the term of this project.

FY 2001

- 1. Improved concentration of measure inequalities such as Talagrand's inequality,
- 2. guarantees on accuracy in terms of assumptions on the distribution generating the empirical process,
- 3. characterizing the computational complexity of basic learning problems within the parameterized complexity framework,
- 4. game theoretic characterizations of feedback effects and route planning in transportation networks.
- efficient algorithms for the class cover type learning problems.

FY 2002

- Continue work on 1-4,
- quantitative understanding of tradeoffs between accuracy and computation resource requirements for Support Vector Machines,
- accuracy guarantees for adaptive methods such as Boosting,
- 5. investigate Markov chain methods to basic learning problems, especially in a distributed game theoretic setting,
- 6. characterize the computational complexity of succinctly specified learning problems in simulation science.

FY 2003

- Continue work on 1-6,
- formal justification and development of accuracy guarantees conditioned on observed statistics of the empirical observations,
- establishment of classical statistical methods such as Maximum Likelihood Estimation and Cross Validation as members of the COLT framework,
- accuracy guarantees for data dependent learning mechanisms,
- characterize the computational complexity of generalized satisfiability problems and their variants as applied to learning models,
- study the game-theoretic/learning aspects of players in deregulated power market and its applications to devising efficient and realistic simulations.
- application to computer intrusion detection problem at LANL,

FY 2004

- Continue work on 1-6,
- alternative learning mechanisms that exploit the trade-off between accuracy and computation,
- application to detection of hard and deeply buried targets,
- accuracy guarantees for randomized algorithms,

Our funding request is for $800K per year for a total of 4 years to be roughly distributed according to the following table.

CIC	TSA	NIS	X	External
240K	240K	90K	80K	150K

References

[1] R. Aumann, M.B. Maschler and R.E. Stearns, *Repeated Games With Incomplete Information*, M.I.T. Press, 1995.

[2] C. Barrett, H.B. Hunt III, M.V. Marathe, S.S. Ravi and D.J. Rosenkrantz, "Computational Aspects of Sequential Dynamical Systems:I," submitted for publication, May 2000.

[3] C. L. Barrett and C. M. Reidys, "Elements of a Theory of Computer Simulation I: Sequential CA over Random Graphs", *Applied Mathematics and Computation*, Vol. 98, 1999, pp. 241–259.

[4] C. L. Barrett, H. S. Mortveit and C. M. Reidys, "Elements of a Theory of Computer Simulation III: Equivalence of SDS", to appear in *Advances in Applied Mathematics*.

[5] P.L. Bartlett, The sample complexity of pattern classification with neural networks: The size of the weights is more important than the size of the network, *IEEE Transactions on Information Theory* **44**, 525–536, 1998.

[6] L. Breiman, Bagging predictors, *Machine Learning*, 26(2), pp. 123-140, 1996.

[7] C. Barrett, R. Jacob and M. Marathe, Formal Language Constrained Path Problems, to appear in *SIAM J. Computing*, Feb. 2000.

[8] R. Carr, S. Doddi, G. Konjevod and M. Marathe, On the Red-Blue Set Cover Problem, in Proc. *11th ACM-SIAM Symposium on Discrete Algorithms (SODA)*, January 2000.

[9] S. Doddi, M.V. Marathe, S.S. Ravi D. Taylor and P. Widmayer, "Approximation algorithms for clustering to minimize the sum of diameters," to appear in Proc. *7th Scandinavian Workshop on Algorithm Theory, (SWAT)*, July, 2000, Bergen, Norway.

[10] D. Fudenberg, D. Levine (Contributor) *The Theory of Learning in Games (Economics Learning and Social Evolution, 2)*, MIT Press, January 1999.

[11] M. Dyer, C.S. Greenhill, L.A. Goldberg, G. Istrate and M. Jerrum The Convergence of the Prisoner's Dilemma Game, submitted to *Artificial Intelligence*

[12] H.B. Hunt III, M.V. Marathe, V. Radhakrishnan, S.S. Ravi, D.J. Rosenkrantz and R.E. Stearns, "NC-Approximation Schemes for NP- and PSPACE-hard Problems for Geometric Graphs," to appear in *J. Algorithms*, (26), pp. 238–274, 1998.

[13] Y. Freund and R.E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. In *Computational Learning Theory: Second European Conference, EuroCOLT '95*, pages 23–37, 1995.

[14] M. Fugate, D. Hush, C. Scovel and R. Christensen, An equivalence relation between parallel calibration and principal component regression, to appear in *Journal of Chemometrics*, 2000.

[15] D.R. Hush, Training a sigmoid node is hard, *Neural Computation*, Vol. 11, pp. 1249-1260, 1999.

[16] D. Hush and C. Scovel, On the VC Dimension of Bounded Margin Classifiers, LA-UR-99-2526., to appear in *Machine Learning*, 2000.

[17] D. Hush and C. Scovel, A New Proof of Concentration of Rademacher Statistics, submitted to *Annals of Probability*, 1999.

[18] D. Hush and C. Scovel, Conditional Performance Bounds for Machine Learning, submitted to *Machine Learning*, 1999.

[19] G. Istrate, M. Marathe and S.S. Ravi, Adversarial models for learning in games, in preparation, June 2000.

[20] M. Kearns and U. Vazirani, *Computational Learning Theory*, International press, November 1998.

[21] M. Muselli, "On convergence properties of pocket algorithm," *IEEE Transactions on Neural Networks*, 8(3), pp. 623-629, 1997.

[22] M.V. Marathe, H.B. Hunt III, R.E. Stearns, and V. Radhakrishnan, "Hierarchically and 1-Dimensional Periodically Specified Generalized CNF Satisfiability Problems with Applications," revised version submitted *J. Computer and System Sciences*, 2000.

[23] N. Nisan, Algorithms for Selfish Agents, in Proc. *16th Annual Symposium on Theoretical Aspects of Computer Science*, Trier, Germany, March 1999. Lecture Notes in Computer Science 1563, pp. 1–15, Springer Verlag.

[24] L. Valiant, A Theory of the Learnable, *Communications of the Association for Computing Machinery (CACM)* **27:11**(1984), 1134–1142.

[25] V. Vapnik *Statistical Learning Theory*, John Wiley, 1998.

5 APPENDIX: Key Participants and Qualifications

The proposed research will be carried out in conjunction with a number of well known scientists at the laboratory and at academic institutions. Specifically, at the laboratory, the scientists span a number of divisions including NIS, CIC, TSA, CNLS and X. External collaborators include distinguished scientists at academic institutions and research laboratories. The PIs have ongoing research collaborations with all of the external participants.

The PIs and the proposed collaborators have been working on various issues related to this proposal. In [16] we have provided the first proof of a result that forms the foundation to Vapnik's Support Vector Machines. In [15] we show that an important learning mechanism in regression and function approximation is computationally intractable. In [17] we provide a new proof of concentration of measure for Rademacher statistics and in [18] we develop and prove the first accuracy guarantees conditioned on statistics of the observed data. In [3, 4, 11, 2] we have characterized the mathematical structure and the computational complexity of several basic combinatorial problems related to a class of discrete dynamical systems. These results have a direct bearing on predicting the outcomes of certain repeated games with incomplete information. In [7], we have characterized the computational complexity and devised efficient algorithms for route finding problems with formal language constraints. These results can be combined with feedback results to obtain efficient learning mechanisms for route discovery in transportation and communication networks. In [8, 9, 12] we have devised efficient algorithms for basic learning problems that can be termed as constrained covering.

The various ingredients of this work will be staffed as follows.

- Generalization error and tradeoffs with computational resources: Hush, Scovel, Koltchinskii, Cannon, Gurvits
- Learning algorithms and computational complexity: Hush, Marathe, Istrate, Howse, Ettinger, Perkins
- Learning in games: Scovel, Stearns, Barrett, Reidys
- Applications: Ettinger, Perkins, Howse, Hush, Marathe

Biographies

Don R. Hush received his Ph.D. in engineering at the University of New Mexico, Albuquerque, New Mexico in 1986. He has served as a technical staff member at Sandia National Laboratory (1986-87), a Professor of Computer Engineering in the EECE Department at the University of New Mexico (1987-98), and a technical staff member at Los Alamos National Laboratory (1998-present). He is a Senior Member of the IEEE and a member of the International Neural Network Society. His research has focused on Machine Learning and Neural Networks for the last 15 years. His research interests include Computational Learning Theory, Machine Learning, Numerical Optimization, Neural Networks, and Signal Processing. He is the author of over 100 scientific publications, a majority of which are in the Machine Learning/Neural Networks area. As of 1997 he had over 220 science citations.

Clint Scovel received his Ph.D. in Mathematics at The Courant Institute of Mathematical Sciences at NYU in 1983 and has been a technical staff member in T-7 from 1986-1989 and in the Computer Research Group CIC-3 since 1989. His research interests include differential

equations, differential and symplectic geometry, symmetry, machine learning, and probability. He was the co-founder of symplectic numerical integration methods for Hamiltonian mechanics. In 1997 his research interests shifted to machine learning and probability, and he became the technical lead on the Medicare Fraud Detection project at LANL. He has 10 scientific publications in data analysis, machine learning, and probability.

Madhav Marathe received his Ph.D. in Computer Science from SUNY, Albany, 1994 under the supervision of Prof. Hunt and Stearns. He Leads the computational research in TSA-2 and the router module in the TRANSIMS project. Expertise in Computational Complexity Theory, Design and Analysis of Algorithms, Simulation Science and Discrete Mathematics. Has published more than 60 research articles in peer reviewed conferences and journals. Pioneered the effort to formulate the notion of approximation algorithms for multi-criteria optimization problems and approximation algorithms for PSPACE and NEXPTIME-hard problems. He is an associate editor of J. of Computing and Information and a recipient of the University Distinguished Dissertation award.

Chris Barrett Group leader TSA-2, Ph.D. Bioinformation Systems, Caltech. Designer and Leader of TRANSIMS project. Research in cognitive science, intelligent control, theoretical and applied issues of very large scale simulations. Certified US Navy Aerospace Experimental Psychologist (cognitive and human factors). Member of National Research Council (TRB) subcommittee for highway and roadway simulation.

Mark Ettinger received his Ph.D. in Mathematics in 1996 from the University of Wisconsin at Madison. He has been at LANL since 1993 and his interests encompass various areas of theoretical computer science including combinatorial games, quantum computation, and algorithmic problems in molecular biology.

James Howse received his Ph.D. in electrical engineering from the University of New Mexico in 1995. He has served as a technical staff member in X-8 at Los Alamos National Laboratory since 1998, and served as a postdoctoral researcher in X-8 from 1996-1998. He has worked for both NASA and Bellcore previously. His research interests include machine learning, control systems, optimization, parameter estimation, time series analysis, and numerical solutions of partial differential equations. He has published papers in the areas of control systems, parameter estimation, and neural networks.

Gabriel Istrate received his Ph.D. in Computer Science from University of Rochester in 1999, and is a Director Funded Postdoctoral fellow in the CNLS and CIC division. His research interests lie in Algorithms and Computational Complexity, with a special interest in problems at the interface with Artificial Intelligence, and probabilistic methods. An ongoing theme of his research has been the study of Phase Transitions in Combinatorial Problems. He recently became interested in theoretical issues related to complex systems, such as (evolutionary) game theory. Recent work in this area involves bounding the time for the emergence of cooperation in a multi-agent system under the Pavlov rule, and the study of learning in games under adversarial models inspired by distributed computing.

Christian Reidys TSA-2, Ph.D. University of Jena, 1994. Expertise in Combinatorics, Graph theory, Probabilistic Methods and Mathematical Biology. Pioneered and leads mathematical research in foundations of simulation science. He has published over 40 papers in peer reviewed conferences and journals.

Simon Perkins received his PhD from the Department of Artificial Intelligence at Edinburgh University in 1998. He is currently a post-doctoral research associate in NIS-2 working on

genetic algorithms and machine learning techniques for remotely sensed image feature classification. His research interests include evolutionary algorithms, machine learning, remote sensing, robotics, and hardware implementations. He has 6 scientific publications in the last two years.

Adam Cannon received his Ph.D. in Applied Mathematics from the Department of Mathematical Sciences at The Johns Hopkins University (2000). His research interests are in statistical pattern recognition, machine learning, applied probability, statistics, and discrete math. His dissertation combined aspects of discrete math, theoretical computer science, and nonparametric statistics to introduce new approximate distance methods for classification and clustering. He is currently a faculty member in the Department of Computer Science at Columbia University in New York. For the summer of 2000 he is visiting the pattern recognition/machine learning team at Los Alamos National Laboratory in New Mexico.

Vladimir Koltchinskii received his Ph.D. in Mathematics at Kiev University in the Ukraine in 1982. He was the head of the Research Laboratory on Computational Statistics in Kiev University up to 1994. In 1992 he was a visiting professor in Mathematics in the University of Connecticut and a visiting scholar in MIT. In 1993-1994 he was a Humboldt Research Fellow at the University of Giessen, Germany. He has been with the Department of Mathematics and Statistics at UNM since 1994 and is currently an Associate Professor and Regents Lecturer there. His research interests are in a broad area of Probability Theory, Analysis, Mathematical Statistics and their applications. He worked on Probability in Banach Spaces, Empirical Processes, Strong Approximation in Probability, Random Matrices, Nonparametric Statistics, Inverse Problems, Multivariate Analysis, Bootstrap, von Mises Calculus. In the last two years his interests shifted to the development of probabilistic methods in machine learning. He has published about 90 articles, 60 of them are research papers in refereed journals and edited volumes. In the last three years he has published 18 articles, many of them in the area of statistical learning theory and its applications to control problems.

Leonid Gurvits received a PhD in Mathematics at Gorky State University in 1985. He is currently at Princeton's Institute for Advanced Studies and has worked at such prestigious institutions as the NEC Research Institute, Technion, the Isaac Newton Institute for Mathematical Sciences in Cambridge, ENS in France, and the Courant Institute of Mathematical Sciences at NYU. He has 71 scientific publications in the fields of Markov Processes, Control Theory and Robotics, Learning Theory, Approximation Theory, Functional Analysis, Matrix Theory and Numerical Linear Algebra.

Richard Stearns received his Ph.D. in Mathematics from Princeton University under the supervision of Prof. Kuhn. He is leading expert in computational complexity theory, algorithms and game theory. He is a distinguished Professor at University of New York at Albany. Along with Juris Hartmanis he has done seminal work in computational complexity theory for which they received the ACM Turing award, the highest research award in Computer Science. He is the fellow of the ACM and along with R.J. Aumann, and M. Maschler has done seminal work in theory of games (repeated games with incomplete information) for which their work received the Lancaster prize.